



### Organización de Datos 75.06, Primer Cuatrimestre de 2019. Examen parcial, primera oportunidad:

Importante: Antes de empezar complete nombre y padrón en el recuadro. Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4) con al menos 20 puntos entre los ejercicios 1 y 2. Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. En el ejercicio 3 elija 2 de los 4 ejercicios y resuelva única y exclusivamente 2 ejercicios. Si tiene dudas o consulta levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen está disponible en forma pública en el grupo de la materia.

"When you play the game of thrones, you win or you die. There is no middle ground." - Cercei Lannister, Game of Thrones

#	1	2	3.1[ ]	3.2[ ]	4	5	6	7	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/10	/10	/15	/15	/10	/10	/100

Nombre:
Padrón:
Corregido por:

1) El servicio meteorológico registra datos del tiempo para todas las ciudades donde posee una base de medición.  Esta información se encuentra en dos RDD. En el primero se tiene información de las bases de medición: (ID_BASE, NOMBRE, PCIA, CIUDAD, LAT, LON). El segundo posee información sobre las mediciones en sf: (TIMESTAMP, ID_BASE, TEMPERATURA, HUMEDAD, PRESIÓN, DIRECCIÓN VIENTO, VELOCIDAD VIENTO). Se desea obtener un reporte para las bases de la Provincia de Buenos Aires. El mismo debe informar los ID y nombre de bases de medición que hayan registrado una variación de temperatura promedio mensual mayor al 30% en el año 2018 (dada la temperatura promedio de un mes, esta se debe comparar con el promedio del mes anterior, para determinar la variación porcentual). Resolver utilizando el API de RDD de PySpark, dando el reporte en un RDD. (***) (15pts)	2) Un importante broker de compra y venta de vehículos online se encuentra dando sus primeros pasos en la preparación de su algoritmo de pricing, es por eso que se encuentra generando algunos features iniciales para experimentar con distintos algoritmos de machine learning.  Para ello cuenta con un archivo con información de todas las transacciones que tuvo en su primer año de operación en el formato (transaction_id, timestamp, vehicle_model_id, price). Por otro lado cuenta con información que fue extrayendo a partir de scrapping durante el último año en el formato (timestamp, source, vehicle_model_id, price). La información puede venir de múltiples fuentes (source), que pueden ser por ejemplo distintos sitios de marketplace o clasificados. Luego de un intenso trabajo previo ha podido unificar los modelos de vehículos que utiliza para sus transacciones con la información que ha podido obtener de otros competidores mediante scrapping. Muchos de los modelos disponibles en la información de scrapping no han sido aún comercializados por la empresa, pero se sabe que se cuenta con precios scrapeados de todos los modelos que se vendieron. Se pide generar utilizando Pandas un dataframe que tenga el siguiente formato (vehicle_model_id, ext_mean_price, ext_std_price, int_mean_price, int_std_price), siendo: - mean_price: precio promedio para ese vehículo. - std_price: desvío estándar del precio para ese vehículo. y los prefijo ext_ y int_ indicando que deben ser calculado sobre respectivamente datos externos (los obtenido vía scraping) y datos internos (los de las transacciones). (***) (15pts)
---	---

- 3) Resolver 2 (dos) y solo 2 de los siguientes ejercicios a elección (si resuelve más de 2 el ejercicio vale 0 puntos, sin excepciones). En cada caso indicar V o F justificando adecuadamente sus respuestas. Si no justifica vale 0 puntos sin excepciones.

a) En SHA-256, es condición necesaria y suficiente que la función de Davis-Meyer sea resistente a colisiones para que la construcción de Merkle-Damgard sea resistente a colisiones (****)	b) Sea $m$ el número de slots de una tabla, y el espacio de claves $U$ es tal que $ U  > (n - 1)m$ , entonces tiene que existir un subconjunto de $U$ de tamaño $n$ que hashea al mismo slot (***)	c) En Cuckoo hashing podemos aumentar el factor de carga máximo antes de rehashear si usamos más de una función de hash. (****)	d) Una reducción de dimensiones no puede generar una distorsión en las distancias menor al error que establece el lema JL. (**) (15pts)
--	--	---	---

4) Sean los puntos $A=(24.21; 0)$ , $B=(0; 225.24)$ y $C=(-14.33; 0)$ . Se quiere construir una familia LSH de forma tal que los puntos a distancia angular menor o igual a $90^\circ$ tengan una probabilidad de colisión de <b>al menos</b> 0.75, y puntos a distancia angular mayor o igual a $135^\circ$ tengan una probabilidad de colisión a <b>lo sumo</b> de 0.4375. Se pide: a) Encontrar cuántos hiperplanos debemos utilizar. ¿Cuál es el valor de $b$ y $r$ ? b) Utilizando los valores de $b$ y $r$ , encontrar hiperplanos que produzcan que A colisione con B; que B colisione con C y que A no colisione con C c) Sea el punto query $Q=(-5; -3)$ . ¿Cuáles puntos serán los candidatos a ser similares? Explique el paso a paso de la consulta. (***) (15 pts)	5) Una planta industrial decidió instalar un sistema monitor de temperatura, a fin de obtener registro de las variaciones que existen, y poder tomar acciones de ser necesario. Dicho monitor cuenta con un sensor que emite cada 5 segundos un registro (fecha: AAAAMMDD, hora: HH:MM:SS, Temperatura: XX.XX, Variación: Numérico, puede ser positivo o negativo). Más allá de las acciones inmediatas que se puedan tomar, esta información se quiere almacenar para realizar consultas o análisis a futuro. Se pide proponer una solución que permita almacenar estos datos comprimidos. Se pueden utilizar uno o más algoritmos de los vistos en clase, o proponer variantes adaptadas a la estructura específica de los datos con los que se cuentan. Se debe explicar cómo queda la estructura final del archivo, y el análisis en el que se basó la solución. (****) (15pts)
---	--

6) Dado los documentos: D1: BATMAN JOKER BATMAN PINGUINO D2: ACERTIJO GORDON JOKER D3: BATMAN ACERTIJO D4: PINGUINO GATUBELA BARBARA a) Construir el índice invertido usando concatenación de términos y utilizando codificación de distancias en tamaño fijo para los punteros de documentos. Indicar el paso a paso en la construcción y todos los archivos que componen el índice final. (4 pts) b) Resolver la consulta de frase "BATMAN PINGUINO", explicando paso a paso. (2 pts) c) Indique qué información agregaría al índice para resolver la consulta de frase más eficientemente. (2 pts) d) Indique para el punto b cuantos accesos se realizaron. (2 pts)	7) Dados los datos del punto 2, diseñar una visualización que permita realizar un análisis de la evolución de precios durante el primer año, y comparar los valores manejados en las transacciones realizadas por el broker, con los valores que manejaron sus competidores. Incluir en la misma visualización algún elemento que permita contrastar el volumen de transacciones manejadas por el broker, contra el total de vehículos de ese modelo ofertados en el mercado. (****) (10 pts)
---	---

1C 2019 - 10P

DATETIME . STBPTIME (COL, F02)

EJ ①

BASES = BASES . FILTER ( LAMBDA X : X[2] == 'BUENOS AIRES' )

BASES = BASES . MAP ( LAMBDA X : ( X[0] , X[1] ) )

BASES = ( ID - BASE , NOMBRE )

MEDICIONES = MEDICIONES . MAP ( LAMBDA X :

( X[1] , X[0] , X[2] ) )

MEDICIONES = ( ID - BASE , TIME , TEMP )

MEDICIONES = MEDICIONES . FILTER ( LAMBDA X :

X[1] > '01.12.2017' AND X[1] ≤ '31-12-2018' )

MEDICIONES = MEDICIONES . MAP ( LAMBDA X :

( X[0] , X[1] . MONTH , X[1] . YEAR , X[2] )  
ID            MES            AÑO            TEMP

MEDICIONES = MEDICIONES . MAP ( LAMBDA X :

( ( X[0] , X[1] , X[2] ) , ( X[2] , 1 ) )  
( ID    MES    AÑO )    TEMP    1

MEDICIONES = MEDICIONES . REDUCEBYKEY ( LAMBDA A, B :

( A[0] + B[0] , A[1] + B[1] ) )

MEDICIONES = MEDICIONES . MAP ( LAMBDA X :

( ( X[0][0] , X[0][1] , X[0][2] ) , ( X[1][0] , X[1][1] , X[1][2] ) )

BASES = ( ID - BASE, NOMBRE )

MEDICIONES = ( ( ID - BASE, MES, AÑO ), PROMEDIO )

A HOJA YO QUIERO QUE MI FILTRO SEA SOLO LAS BASES  
QUE EL PROMEDIO DEL MES ACTUAL CON EL ANTERIOR  
SEA MAYOR AL MES ANTERIOR

PROMEDIO ACTUAL > 0,30 \* PROMEDIO MES  
ANTERIOR

MEDICIONES MES ANTERIOR = MEDICIONES . MAP ( LAMBDA X :  
( ( X[0][0], ( X[0][1] + 1 ), X[0][2] ), PROMEDIO ) )

MEDICIONES ANTERIORES = MEDICIONES . MAP ( LAMBDA X :  
( ( X[0][0], 1, 2018 ), X[1] ) IF X[0][1] = 12  
AND X[0][2] = 2017 ELSE X ) )

MEDICIONESANT = MEDICIONESANT . FILTER ( LAMBDA X :  
X[0][1] < 13 )

EJ 2

INTERNAS = ( TRANSACTION-ID, TIME, VEHICLE-ID, \$ )  
0 1 2 3

EXTERNAS = ( TIME, SOURCE, VEHICLE-ID, \$ )  
0 1 2 3

CREAR UN DF

( VEHICLE-ID, EXT-MEAN-PRICE, EXT-STD-PRICE,  
INT-MEAN-PRICE, INT-STD-PRICE )

INT = INTERNAS. GROUPBY ([ 'VEHICLE-ID' ]). AGG (  
| PRICE : [ 'MEAN', 'STD' ] | ). RESET-INDEX

INT. COLUMNS = [ 'VEHICLE-MODEL-ID', 'INT-MEAN-PRICE'  
INT-STD-PRICE ]

EXT = EXTERNAS. GROUPBY ([ 'VEHICLE-ID' ]). AGG (  
| PRICE : [ 'MEAN', 'STD' ] | ). RESET-

EXT. COLUMNS = [ '---' ]

RESULTADO = INT. MERGE ( EXT, HOW = , IN = VEH )

EJ ④

$$A = (24.21, 0) \quad C = (-14.33, 0)$$

$$B = (0, 225.24)$$

- LOS PUNTOS DE  $\leq 90^\circ \quad P_C \geq 0,75$
- LOS PUNTOS DE  $\geq 135^\circ \quad P_C \leq 0,4375$

† A)  $P_{B_1, B_1} = 1 - (1 - (1 - 0,5^\circ)^R)^B \geq 0,75$

$$P_{B_1, B_2} = 1 - (1 - (1 - 0,75^\circ)^R)^B \leq 0,4375$$

LAS DISTANCIAS EN GRADOS LAS PASO A DECIMAL

$$\text{de}_1 = 90^\circ \rightarrow \frac{90}{180} = 0,5$$

$$\text{de}_2 = 135^\circ \rightarrow \frac{135}{180} = 0,75$$

CON DESMOS OBTENGO  $B=2 \quad R=1$

$\downarrow$   
CANTIDAD  
DE TABLAS

→ FUNCIONES  
DE MH POR  
TABLA

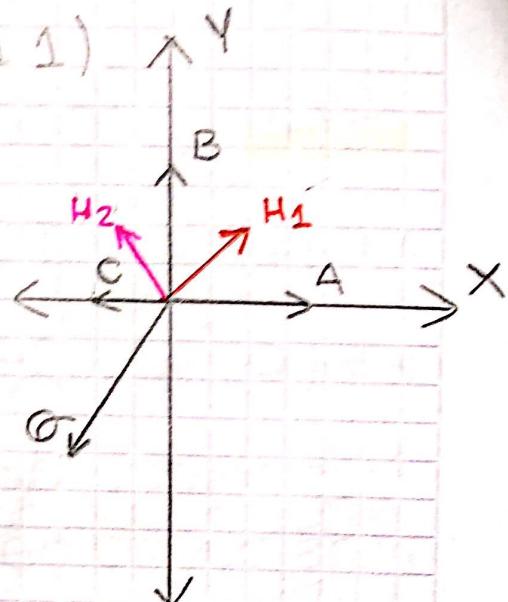
COMO  $B * R = 2$  NECESITO 2 HIPERPLANOS  
PARA PODER REPRESENTAR  
LOS 2 MH PARA EL PROCESO  
DE LSH.

B) PARA  $B=2$   $B=1$

$$H_1 = (1 \ 1)$$

$$H_2 = (-1 \ 1)$$

- A COLISIONE CON B
- B COLISIONE CON C
- A NO COLISIONE CON C



	A	B	C
H1	1	1	0
H2	0	1	1

$B_1$   
 $B_2$

(REEMPLAZO LOS )  
-1 POR 0

CUANDO COLISIONAN  
TIENEN  $P.I > 0$  Y CUANDO  
NO TIENEN  $P.I < 0$

0	1
C	A, B

$B_1$

0	1
A	B, C

$B_2$

COLISIONAN



SI EL ANGULO ENTRE DOS VECTORES  $< 90^\circ \rightarrow P.I (+)$

(ÁREAS DE INTERSECCIÓN  
COMO ME ENSEÑÓ TOMI)

$> 90^\circ \rightarrow P.I (-)$



NO COLISIONAN

C)  $Q=(-5 \ -3)$

$$(-5 \ -3) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = -8 = 0$$

$$(-5 \ -3) \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} = 2 = 1$$

CANDIDATOS

	A	B	C	Q
H1	1	1	0	0
H2	0	1	1	1

C, Q	A, B
------	------

$$Q_1 = \{C\}$$

ORDOC  $\{B, C\}$

	A	B, C
		Q

$$Q_2 = \{B, C\}$$

NOTA

EJ (6)

DOC 1: BATMAN JOKEZ BATMAN PINGUINO

DOC 2: ACERTIJO GORDON JOKEZ

DOC 3: BATMAN ACERTIJO

DOC 4: PINGUINO GATUBELA BARBADA

(3) (6) (7) (5)  
 BATMAN JOKEZ PINGUINO ACERTIJO GORDON  
 (1)  
 GATUBELA BARBADA  
 (4) (2)

A)

- |           |   |          |   |
|-----------|---|----------|---|
| BATMAN    | 1 | ACERTIJO | 2 |
| \JOKER    | 1 | ACERTIJO | 3 |
| \BATMAN   | 1 | BARBADA  | 4 |
| PINGUINO  | 1 | BATMAN   | 1 |
| \ACERTIJO | 2 | BATMAN   | 1 |
| \GORDON   | 2 | BATMAN   | 3 |
| \JOKER    | 2 | GATUBELA | 4 |
| \BATMAN   | 3 | GORDON   | 2 |
| \ACERTIJO | 3 | JOKER    | 1 |
| PINGUINO  | 4 | JOKER    | 2 |
| \GATUBELA | 4 | PINGUINO | 1 |
| \BARBADA  | 4 | PINGUINO | 4 |

ACERTIJO 2,3

BABBADA 4

BATMAN 1,3

GATUBELA 4

GOBDON 2

JOKER 1,2

PINGUINO 1,4

ACERTIJO 2,1

BABBADA 4

BATMAN 1,2

GATUBELA 4

GOBDON 2

JOKER 1,1

PINGUINO 1,3

LOS DOCUMENTOS QUE TIENEN EN LA MANO FIGURA  
S=0 DE 8 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44

ACERTIJO BABBADA BATMAN GATUBELA GOBDON JOKER PINGUINO  
0 8 15 21 29 35 40

2 1 4 1 2 4 2 1 1 1 3  
0 2 3 5 6 7 8 9

0100 11 0001 11 01 00 00  
4 6 10 12 14

LEXICO

PUNTEDOS

	LEXICO	PUNTEDOS	
ACERTIJO	0	0	0010
BABBADA	8	4	18
BATMAN	15	6	
GATUBELA	21	10	
GOBDON	29	12	
JOKER	35	14	
PINGUINO	40	18	

C) PARA QUE SEA MAS EFICIENTE AGREGARIA LA FRECUENCIA DEL TERMINO EN CADA DOC Y LA POSICIÓN, COMO HACEMOS CON CONSULTAS DE PROXIMA FRECUENCIA YA QUE LA FIDASE ES UNA CONSULTA DE PROXIMA FRECUENCIA CON D=1.

B) Q = "BATMAN PINGUINO"

DESOLVEMOS LA CONSULTA CON BUSQUEDA BINADIA

LONG( INDICE ) = 7

MITAD → POS 3

RESUELVO PRIMERO "BATMAN"

POS 3 → ( 1 INDICE , 1 DISCO )

PARA ACCEDER A LA TABLA Y 1 D

PARA IR AL CONCATENADO Y VER QUE

HAY

GATUBELA > BATMAN ?

SI !

VUELVO A APLICAR --

ME QUEDO CON LA PRIMERA MITAD

LONG( IN ) = 3 → MITAD → POS = 1

POS 1 → BARBADA ( 1 INDICE - 1 DISCO )

1 INDICE PARA VER QUE HAY EN  
TABLA - 1 DISCO PARA IR AL CONCT  
Y VER QUE HAY .

• 5

→

→

BARBADA > BATMAN

NO !

POS 2 → BATMAN ( 1 DISCO - 1 INDICE )

ES LO QUE BUSCO : 1 DISCO

POS 3 → 12 → DOC 1 DOC 3

AHORA DESUELVO "PINGUNO"

LONG TABLA = 7 → MITAD 3

POS 3 → GATUBELA (YA TENIA SUS ACCESOS)

GATUBELA ≥ PINGUNO

NO

ME QUEDO CON LA MITAD INFED1012

LONG 3 → MITAD = POS 5

POS 5 → JOKER (1 INDICE - 1 DISCO)

JOKER ≥ PINGUNO?

NO!

POS 6 → PINGUNO

(1 INDICE - 1 DISCO)

ES LO QUE BUSCO! !

1 DISCO PARA IR AL  
CONCATENADO DE DOCUMENTOS

POS 9 → 13

DOC 1, DOC 4

Y ENTRAR AMBAS

DOC 13

+ 1 PARA VERIFICAR  
SI POSTA ES ASI

(( 5 INDICES - 7 DISCOS ))